Taxonómia felismerése dokumentumszerkezetből

Lendvai Piroska

Tilburg University, Dept. of Language and Information Science
PO Box 90153, 5000 LE, Tilburg, Hollandia

p.lendvai@uvt.nl

Kivonat: Munkánk orvosi enciklopédiák szövegéből kinyerhető taxonomikus kapcsolatok automatikus felfedezésére irányul, melyet szabadszavas, egészségügyi tematikájú kérdések automatikus megválaszolásában használunk fel. Az enciklopédiák szócikkeit különböző szövegszegmensi szinteken a témakörre jellemző szemantikai annotációval láttuk el. Mesterséges intelligencia alapú tanulási kísérleteket írunk le, amelyek során a taxonomikus kapcsolatok automatikus felismerésének betanítása és értékelése történik.

1 Bevezetés

A holland ROLAQUAD projekt keretében fejlesztett intelligens válaszadó rendszer célja, hogy szabadszavas, egészségügyi tematikájú kérdéseket válaszoljon meg. A rendszer alapját két, holland nyelvű orvosi enciklopédia kézzel annotált szócikkei képezik. A rendszer felismeri a felhasználó kérdésében a kérdezett tárgyszót (pl. „agyhártyagyulladás"), s hogy annak mely aspektusára kérdez rá a felhasználó (pl. „tünetei"). Ezután a referenciadokumentumok szemantikai annotációjához illeszti ezeket, majd a legpontosabban illesztett dokumentumrészt adja vissza válaszként.

Előfordulhat azonban, hogy a felhasználó kérdése alulspecifikált, például mert a kérdezőnek nincsenek pontos ismeretei az adott területről. Ilyen kérdés lehet a „Mik az agyhártyagyulladás tünetei?", mert a rendszer referenciaszövegében az „Agyhártyagyulladás" szócikk két szakaszában is előfordul a `'Tünetek'` szemantikai annotáció. A helyes válaszadáshoz szükséges felismerni, hogy a szócikk az agyhártyagyulladás két típusát is körülírja, vagyis taxonomikus kapcsolatokat tartalmaz, és hogy emiatt a felhasználót a kérdése pontosítására kell megkérni.

Ahhoz, hogy a rendszer dinamikus módon tudjon ilyesfajta visszakérdezéseket generálni, szükséges, hogy a referenciadokumentumokból automatikusan ki tudja szűrni azokat, amelyek a címszóban megnevezett entitásnak több altípusával is foglalkoznak. Munkánk erre tesz kísérletet, a dokumentum szerkezetére vonatkozó szemantikai annotáció alapján. Az alkalmazott tanuló algoritmusnak azt kell felismernie, hogy a címszóban megjelölt entitás altípusait tárgyalja-e az adott enciklopédia-szócikk olyan részletesen, hogy a címszóban megjelölt entitás vagy annak egy aspektusa végeredményben az altípusok által definiálódik. Pl.: kortikoszteroidok{alkalmazása külsőleg;alkalmazása belsőleg}, vékonybél-daganat{jóindulatú;rosszindulatú}, steri-

lizálás{férfiaknál;nőknél}, stb. A feladatot kétfajta megközelítésben is elvégezzük. Ezekben az algoritmus a szócikkek különböző jellemzőit használja fel a tanulás során, pl. az abban előforduló szavakat, egészségügyi fogalmakat, statisztikai gyakoriságot, stb.

A javasolt eljárás nem morfológiai/szintaktikai alapú [2], hanem közvetlenül a dokumentumok szerkezete és az azok fölötti szemantikai tartalmak alapján azonosítja a taxonomikus kapcsolatot. Korpuszunk dokumentumai kevesebb strukturális hierarchiát mutatnak, mint a [4] által ontológia létrehozásához felhasznált szövegek, a klasszifikáció pedig nem szegmentálásra [1], hanem előre meghatározott szövegszegmensek közötti taxonomikus kapcsolatok felfedezésére irányul.

A következőkben bemutatjuk a felhasznált korpusz szemantikai annotációjának elvét és a különböző szemantikai címkéket. A 3. szakasz a konkrét gépi tanulási kísérleteket írja le, részletezve az alkalmazott algoritmust, a két tanulási feladatot, a tanulásban felhasznált attribútumokat, és a kapott eredményeket. Az utolsó részben összefoglaljuk és értékeljuk munkánkat.

2 A korpusz szemantikai annotációja

A rendszer által felhasznált referencia-dokumentumgyűjtemény a holland nyelvű Merck orvosi kézikönyv és a Spectrum egészségügyi enciklopédia szócikkeiből áll. Korlátozott számú, a témakörre jellemző szemantikai annotációt kaptak a szavak szintjén a fogalmak, a mondatok szintjén a mondat témája, a szakaszok szintjén pedig a szakasz témája. Pl. az "Agyhártyagyulladás" szócikk fordításának második mondata:

<SZAKASZ: `Definíció;Ok`> ... <MONDAT: `Definiál;Fertőz;Okoz`> *A betegség kórokozói különféle* <FOGALOM: `mikroorganizmus`>*vírusok*</FOGALOM> *és* <FOGALOM: `mikroorganizmus`>*baktériumok*</FOGALOM> *lehetnek*.</MONDAT> ... </SZAKASZ>

A teljes korpusz több, mint 3000 dokumentumból áll, ezek 54%-a azonban nem használható fel a kísérletekben, mert szerkezetileg csak egyetlen szakaszból állnak. A több szakaszból álló dokumentumok között 128 olyan található, amely szemantikailag rekurzív szerkezettel rendelkezik, vagyis a bevezető szakaszt követően legalább két olyan szakasza van, amelyek temájukban megegyeznek, pl. két szakasz is tárgyal `Tünetek`-et vagy `Megelőzés`-t.[21]

A kézi annotálás a következő protokoll alapján történt. Egy dokumentum szakaszához 15 különböző címkét lehet hozzárendelni, pl. `Definíció` ('a szakasz a címszó-entitás definícióját tartalmazza'), `Ok` ('a szakasz egy entitás előfordulásának okát írja le'), `Megelőzés`, stb. A teljes címkelistát az 1. Táblázat első oszlopa mutatja. Egy szakaszhoz természetesen több címke is hozzárendelhető.

[21] A kísérlet annotálatlan szövegeken is elvégezhető, ha azok konzisztens (al-)alcímeket tartalmaznak.

Szakasztípus	Mondattéma	Fogalom
alkalmazás	jellemez	testi funkció
ok	okoz	testrész
következmény	alfaja	betegség
fertőzés	fertőz	betegség jellemzője
definíció	definiál	betegség tünete
diagnózis	diagnosztizál	diagnosztikai eljárás
betegségek	hasonlít	időtartam
elsősegély	szinonima	mikroorganizmus
előfordulás	előfordul	személy
mellékhatások	mellékhatása	személy jellemzője
kezelés	kezel	kezelés
tünetek	tünete	kezelés jellemzője
megelőzés	megelőz	
módozatok		
formák		

1. Táblázat. Szemantikai annotáció címkéi a dokumentum három szintjén.

Mondatszinten 13 témát annotáltunk, egy-egy mondatot szintén több címke is jellemezhet; ezeket lásd az 1. Táblázat középső oszlopában. A szavak illetve a szókapcsolatok szintjén 12 egészségügyi fogalomtípust címkéztünk; lásd az 1. Táblázat harmadik oszlopát.[22]

3 Gépi tanulási kísérletek

A kísérleteket felügyelt tanulási feladatként formalizáljuk, melyeket a TiMBL szoftvercsomag 5.1 verziójának IB1 algoritmusával végzünk el[23]. Az algoritmus a *k*-legközelebbi szomszéd ('*k*-nearest neighbour', '*k*-NN') tanulási módszert használja, lásd pl. [5]. Ennek a felügyelt módszernek a működési elve példányalapú tanulás, vagyis egy feladatot példák attribútum-vektoraként jelenítünk meg, és az algoritmus ezekhez tanul meg osztályokat rendelni. Az algoritmust alapbeállításokkal futtattuk (*k*=1, a példák között euklidészi távolság mérése, 'gain ratio' attribútumsúlyozás). A kísérleteket a „kihagyok egyet" ('leave-one-out') predikció módszerével folytattuk le. Két kísérletsorozatot végeztünk el, amelyekben különbözőképpen közelítettük meg a taxonomikus kapcsolatok feltárását.

[22] A domén-entitások jó része egyszerű, formai jellemzők alapján kinyerhető annotálatlan szövegekből is, erről lásd pl. [3].

[23] Daelemans, W., Zavrel, J., Van der Sloot, K., and Van den Bosch, A. (2004). TiMBL: Tilburg Memory Based Learner, version 5.1, Reference Guide. ILK Technical Report Series 04-02. http://ilk.uvt.nl/timbl

3.1 Osztályozási feladatok

Az első kísérletsorozatban az algoritmus feladata annak eldöntése, hogy egy dokumentum két adott szakasza taxonomikus testvérpárt ír le vagy sem. Az „Agyhártyagyulladás" szócikk például négy szakaszból áll: (1) bevezetés, (2) „Bakteriális agyhártyagyulladás", (3) „Megelőzés", és (4) „Vírusos agyhártyagyulladás". Ebből a {2,4} szakaszok taxonomikus testvérpárt alkotnak: azonos rangú taxonomikus relációban állnak a szócikk által leírt entitással (agyhártyagyulladás), mivel mind a (2), mind a (4) szakasz témája a szemantikai annotáció szerint `'Okoz'`, `'Tünetek'`, és `'Kezelés'`. A feladat itt annak a felismerése, hogy a két szakasz tartalmi átfedéseket tartalmaz ugyan, de a témák fő argumentuma különbözik egymástól. Az algoritmusnak tehát nemcsak a két szakasz közötti hasonlóságokat, de a különbségeket is számon kell tudni tartani.

A második kísérletsorozatban a feladat olyan pozitív példák felismerése, ahol a szakaszpár egyik tagja a címszóban megjelölt egészségügyi fogalmat általánosságban jellemzi, míg a másik annak altípusát írja le. A példadokumentumban az {1,2} és az {1,4} szakaszpárok írnak le ilyen, alárendeltségi kapcsolatot egy általános fogalom és annak alfaja között, mert a bevezető az általános fogalmat, az agyhártyagyulladást írja körül, míg a 2. illetve a 4. szakasz annak egy specifikus alfaját. Ez a feladat látszólag még nehezebb, mint az első megközelítés, mert egy enciklopédia-szócikk bevezetője tartalmilag szükségszerűen utal az összes következő szakaszra, és a szakaszok is utalhatnak egymásra – az algoritmus dolga itt az, hogy felismerje, hogy az egyik szövegszegmens a másik egy adott elemét részleteiben tárja fel. Bizonyos szempontból a szegmensek közötti kapcsolatot nemcsak alárendeltséginek foghatjuk fel, de anaforikusnak is.

Az algoritmus számára az {1,3} szakaszpár mind a taxonomikus testvérpárnak, mind az alárendeltségi kapcsolatnak negatív példája.

Mivel a dokumentumgyűjteményben két különböző típusú enciklopédia szócikkei szerepelnek, hasznosnak láttuk ezeket egymástól különválasztva feldolgozni. A Spectrum enciklopédia szócikkei igen következetesen strukturáltak, a szakaszok címei konzisztensen visszatérnek, vagyis szerkezetelemzéshez „ideális" anyagot nyújtanak. A Merck kézikönyv dokumentumaiban a szerkezet lazább, az alcímek esetlegesebbek, a szócikkek pedig hosszabbak, mint a Spectrumban, ezért a Merck feldolgozása inkább hasonlítható egy „valós" szemantikai elemzési környezethez.

A két különböző osztályozási feladatban szükségszerűen különbözik a vonatkozó pozitív és negatív példák száma is. A taxonomikus testvérpárok feladathoz 174 pozitív és 523 negatív példát tudunk generálni a Spectrum enciklopédiából, és jóval kevesebbet a Merck kézikönyvből (49 pozitív, 161 negatív példa). Az alárendeltségi kapcsolat meghatározásának feladatához valamivel több pozitív és valamivel kevesebb negatív példa áll rendelkezésre, ami elősegítheti a hatékonyabb osztályozást (Spectrum: 255 pozitív és 442 negatív, Merck: 51 pozitív és 159 negatív példa).

3.2 Felhasznált attribútumok

A szakaszpárokat különbözőképpen jelenítjük meg az egyes kísérletek során. Az attribútumvektor komponensei numerikus elemekből (főként bináris bitekből) állnak, amelyek a következő információt hordozzák: a két szakaszban előforduló

(a) közös szavak (szóhalmazban, 'bag-of-words')
(b) közös szóhármasok (trigram-ok)
(c) dokumentumcím – szakasz alcím(ek) – vizsgált szakasz(ok) közös szavai
(d) közös egészségügyi fogalmak
(e) közös mondattémák.

Fontos tudni, hogy egy-egy attribútumcsoport kódolása nagyságrendekkel különbözhet egymástól: a szóhalmaz vektora pl. 7288 elemből áll, mert ekkora a korpusz lexikonja. Ha egy szó a vizsgált szakaszok mindegyikében előfordul, a szót jelző bit értéke 2, ha csak az egyik szakaszban, a bit értéke 1, ha egyik szövegszegmensben sem fordul elő, a bit értéke 0. A szóhármasok vektora 1155 elemből áll, mert ekkora a korpuszban a három vagy annál nagyobb (jelen esetben: 36-ig terjedő) gyakorisággal előforduló trigramok lexikonja. A dokumentumcím-szakaszalcím(ek) egybeesése viszont mindössze 4 bitből áll, a közös fogalmaké 12, a közös mondattémáké 13 elemből (lásd 1. Táblázat).

3.3 Eredmények

Az algoritmus teljesítményét többféle mérték szerint is értékeltük: globálisan számított mértékek a pontosság ('accuracy', az általános hibaszázalék ellentettje), a mikro-F-pontszám (az összes példa alapján kiszámított F), a makro-F-pontszám (a két osztály alapján kiszámított F), valamint az osztályokra levetített pontosság ('precision'), teljesség ('recall'), és ezek harmonikus középértéke, az F-pontszám (2PreRec/Pre+Rec). Az értékelés során a legnagyobb figyelmet a pozitív példák klasszifikációjára vonatkozó F-pontszámnak szenteljük, mert ez mutatja, mennyire jól képes az algoritmus a fogalmi taxonómia különböző elemeinek (mellérendelt kapcsolatban lévő „testvéreknek", vagy „alá-fölérendelő" hiperonim-hiponim kapcsolatoknak) a felismerésére.

Korpusz	Attribútum				+ osztály			– osztály		
		Acc	Fmik	Fmak	Pre	Rec	F	Pre	Rec	F
Spectrum	szóhalmaz	55	56	44	20	23	20	72	66	69
	szóhármas	*61*	*61*	*48*	*22*	*21*	*21*	*74*	*75*	*74*
	(al-)címek	86	85	78	98	47	64	85	100	92
	fogalmak	*75*	*75*	*67*	*51*	*49*	*50*	*83*	*84*	*84*
	mondattémák	88	88	85	78	76	77	92	93	92
Merck	szóhalmaz	73	74	65	44	57	50	86	78	81
	szóhármas	*71*	*71*	*58*	*37*	*35*	*36*	*80*	*82*	*81*
	(al-)címek	79	75	61	69	22	34	80	97	88
	fogalmak	*69*	*69*	*56*	*33*	*33*	*33*	*80*	*80*	*80*
	mondattémák	79	80	73	55	65	60	89	84	86

2. Táblázat. Mellérendelt viszonyú taxonomikus testvérpárok meghatározása a kétfajta korpuszban, különböző attribútumok alapján.

Az első kísérletsorozat eredményeit a 2. Táblázat tartalmazza. Megállapítható, hogy legjobb eredményt akkor tudtuk elérni taxonomikus testvérek azonosításában, ha a szakaszpárokat az azokban előforduló azonos mondattémákként ábrázoltuk: a szabadabb formátumú Merck szövegeiben 60 F-pontszámot, a Spectrum szövegeiben pedig, amelynek dokumentumai szabályosabb szerkezetbe rendezettek, 77 F-pontszámot értünk el. A Spectrum anyagán a második legmagasabb F-pontszámot (64) a dokumentumcím – szakasz alcímek – közös dokumentumszavak egybeesésének információja alapján zajló kísérletben értük el. Ebből arra következtetünk, hogy dokumentumszerkezet alapján szemantikai tartalmat fel lehet ismerni abban az esetben, ha a szerkezet jelölése következetes. A szakaszokban szereplő egészségügyi fogalmak csak harmadrangú információt nyújtanak arról, hogy adott szócikk két szegmense tartalmilag egymás mellé rendelhető-e.

A Merck kézikönyv szócikkein elért eredményekből kitűnik, hogy ezeknek a dokumentumoknak a felépítése más, mint a Spectrumban, mert az (al-)címek egybeesésének információja a pozitív osztályt nem, a negatív osztályt viszont igen jól képes jellemezni (88 F). Megállapítható, hogy az azonos fogalmi körbe tartozó, de különböző séma alapján felépített dokumentumokban más és más attribútumcsoportok hordoznak taxonómiai információt. A leginformatívabb természetesen az, hogy mely mondattémák esnek egybe a két szegmens között; ezt optimális esetben a témákkal egybeeső alcímek jelzik.

A szóhalmaz, illetve a szóhármasok által hordozott információ a Merck anyagán jobb eredményt ad, mint a Spectrumén, ami valószínűleg azzal magyarázható, hogy a Merck szócikkei hosszabbak és szabadabb megfogalmazással íródtak. Ez utóbbira tanú az is, hogy a Merckben a témaköri fogalmak megléte, illetve valószínűleg inkább azoknak a hiánya, kevesebb információt tud nyújtani, mint maguk a dokumentumban szereplő szavak (33 F, lásd a táblázat utolsó előtti sorát). A szövegekben megjelenő fogalmak statisztikailag tulajdonképpen csak egy esetben adnak jobb eredményt, mint akár a szóhalmaz, akár az alcímek egybeesése: az alá-fölérendeltségi kapcsolat megállapításakor a Spectrum anyagán (51 F). Ezzel rá is tértünk a második kísérletsorozat tárgyalására (lásd: 3. Táblázat).

Korpusz	Attribútum				+ osztály			– osztály		
		Acc	Fmik	Fmak	Pre	Rec	F	Pre	Rec	F
Spectrum	szóhalmaz	55	54	49	36	28	31	63	71	67
	szóhármas	*54*	*53*	*48*	*35*	*28*	*31*	*62*	*69*	*66*
	(al-)címek	69	64	59	70	27	39	69	93	79
	fogalmak	*64*	*64*	*61*	*51*	*51*	*51*	*72*	*71*	*71*
	mondattémák	85	85	84	78	83	80	90	87	88
Merck	szóhalmaz	79	77	68	58	43	49	83	90	84
	szóhármas	*74*	*74*	*64*	*46*	*45*	*45*	*82*	*83*	*83*
	(al-)címek	76	65	43	-	-	-	75	100	86
	fogalmak	*78*	*77*	*69*	*56*	*49*	*52*	*84*	*87*	*86*
	mondattémák	83	82	74	69	53	60	86	92	89

3. Táblázat. Alá-fölérendeltségi viszonyú (hiperonim-hiponim) szakaszpárok meghatározása a kétfajta korpuszban, különböző attribútumok alapján.

Érdekes megfigyelni, hogy bár a hiperonim-hiponim kapcsolat meghatározása nehezebb feladat lehet, többek között mivel a rövidke bevezető szakasz anyagára kell támaszkodni, aminek nincsenek alcímei, de a korábban tárgyalt anaforikus jelleg miatt is, a 3. Táblázat pontszámai mégis némileg magasabbak és kiegyensúlyozottabbak, mint a mellérendeltségi feladaton elértek. Technikai kérdés, hogy ez vajon annak köszönhető-e, hogy ebben a feladatban valamennyivel több pozitív példa raktározható el a memóriában a tanulási fázis során.

Fontos eredmény, hogy legmagasabb pontszámot ebben a modellben szintén a mondattémák közötti átfedés alapján lehet elérni: a Merck szövegekben 60 F-pontszámot (ez megegyezik a taxonómiának testvérpárok alapján történő felismerésével), a Spectrum szövegeiben pedig 80 F-pontszámot értünk el, ami magasabb, mint a testvérpárok alapján történő felismerés esetében.

Természetesen a szakasz alcímek egybeesése ehhez a feladathoz nem adhat plusz információt, mert a bevezető szakasznak, ami mindig a szócikk első szegmense, soha nincs alcíme. A szakaszokban szereplő egészségügyi fogalmak a Spectrum esetében ismét viszonylag jól jellemzik, hogy adott szócikk két szegmense tartalmilag egymásra mutat egy alá-fölérendeltségi kapcsolatban, a Merck anyagán viszont gyakorlatilag nem adnak többletinformációt az egyszerű (bár nagy számú) szóhalmazhoz képes.

4 Értékelés

Munkákban arra tettünk javaslatot, hogyan lehet gépi tanulási kísérleteket felépíteni fogalmi taxonómia elemeinek kinyerésére strukturált, szemantikailag annotált dokumentumokból, jelen esetben holland, egészségügyi témájú enciklopédia-szócikkekből. A kísérleteket az motiválja, hogy olyan általános módszert találjunk, amelyet következetesen felépített, leíró jellegű dokumentumokra – pl. enciklopédiák, wikipédiák, értelmező szótárak – lehet alkalmazni taxonómia kinyerésére. Megállapítottuk, hogy a taxonómia komponenseit legalább kétféle modellel írhatjuk le: kereshetjük az egy dokumentumban előforduló taxonomikus testvérpárokat, illetve közvetlenül az általános fogalmat és annak egy altípusát. A kísérletekhez példaalapú tanuló algoritmust használtunk, amelynek betanítása öt különböző attribútumcsoporton történt. Mindkét módszerrel megközelítőleg azonos eredményt értünk el, a legmagasabb F-pontszámot (80) a Spectrum egészségügyi enciklopédiából generált példákon: az algoritmus egy általános egészségügyi fogalmat és annak egy altípusát leíró szakaszpárokat azonosított be a szakaszok tematikai egybeesése alapján. Ez első hallásra triviálisnak tűnhet, azonban egyáltalán nem kézenfekvő, hogy a tematikai egybeesés éppen alá-fölérendeltségi kapcsolatra utal, hiszen éppúgy utalhat egy általános anafora-katafora vagy rész-egész kapcsolatra is, hiszen egy dokumentum bevezető szakaszának funkciója, hogy a teljes mondanivalót előrevetítse. Ezért a modellben a negatív példák felismerését szintén nagy pontossággal kell megoldani. A táblázatokból látható, hogy a negatív példák osztályozása jó eredménnyel történik.

A tematikai egybeesés attribútumvektort kézzel annotált címkékből generáltuk. A jövőben arra fogunk sort keríteni, hogy ezt az attribútumot gépi tanulással ki tudjuk nyerni a mondatokból, és további, a szövegről magas szintű szemantikai és morfo-szintaktikai információt közvetítő attribútumokkal egészítsük ki.

A viszonylag kevés számú példa és a korpusz „zajossága” – nem szakértők által történt annotálása és szócikk-szegmentálása – valószínűvé teszi, hogy az itt bemutatottaknál egy színvonalasabban felcímkézett korpuszon jobb eredményeket lehetne elérni a javasolt módszerrel. Amennyiben a taxonomikus kapcsolatokat megbízhatóan tudjuk felismerni, a folyamatot beépítjük az orvosi válaszadó rendszerbe, a taxonómia elemeit pedig ontológia létrehozására használjuk fel.

Bibliográfia

1. Cho, P., Taira, R., Kangarloo, H.: Automatic Segmentation of Medical Reports. Proc. of AMIA Symposium (2003) 155-159
2. Cimiano, P., Pivk, A., Schmidt-Thieme, L., Staab, S.: Learning Taxonomic Relations from Heterogeneous Sources of Evidence. In: Buitelaar, P., Magnini, B., Cimiano, P. (Eds): Ontology Learning from Text: Methods, Applications, Evaluation. IOS Verlag (2005)
3. Lendvai, P.: Conceptual Taxonomy Identification in Medical Documents. In: Proc. of The Second International Workshop on Knowledge Discovery and Ontologies (2005) 31-38
4. Makagonov, P., Figueroa, A., Sboychakov, K., Gelbukh, A.: Learning a Domain Ontology from Hierarchically Structured Texts. Proc. of ICML workshop on Learning and Extending Lexical Ontologies by using Machine Learning Methods (2005) 50-57
5. Mesterséges Intelligencia. Szerk.: Futó, I. Aula Kiadó (1999)